CAN DEEPFAKE SPEECH BE RELIABLY DETECTED?

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in text-to-speech (TTS) systems, particularly those with voice cloning capabilities, have made voice impersonation readily accessible, raising ethical and legal concerns due to potential misuse for malicious activities like misinformation campaigns and fraud. While synthetic speech detectors (SSDs) exist to combat this, they are vulnerable to "test domain shift", exhibiting decreased performance when audio is altered through transcoding, playback, or background noise. This vulnerability is further exacerbated by deliberate manipulation of synthetic speech aimed at deceiving detectors. This work presents the first systematic study of such active malicious attacks against state-of-the-art open-source SSDs. White-box attacks, black-box attacks, and their transferability are studied from both attack effectiveness and stealthiness, using both hardcoded metrics and human ratings. The results highlight the urgent need for more robust detection methods in the face of evolving adversarial threats.

1 INTRODUCTION

Recent years have witnessed a remarkable advance in text-to-speech (TTS) systems (Kreuk et al., 2022; Borsos et al., 2023; Leng et al., 2023; Saeki et al., 2023; Shen et al., 2023; Wang et al., 2023; ChatTTS, 2024; Chen et al., 2024; xTTS, 2024; ElevenLabs, 2024; Le et al., 2024; Lux et al., 2024). Some of these systems possess the zero-shot / few-shot voice cloning capability (Biadsy et al., 2024; Cooper et al., 2020; Casanova et al., 2022; Ye et al., 2024; Le et al., 2024), that can mimic someone's voice using only a brief sample of that person's speech recording. The widespread access of these systems, either through open-source projects or commercial APIs, has made voice impersonation easier than ever. This raises significant ethical and legal concerns as the capability can be easily misused for misinformation campaign, fraud, copyright infringement, *etc*. As an instance, a scammer utilized synthetic audio to mimic President Biden in unlawful robocalls during a New Hampshire primary election, resulting in a $6 million penalty and felony accusations (Coldewey, 2024). Besides, numerous instances of synthetic audio misuses can be found online if searching with "deepfake audio", which underscores the urgent need to address this growing problem.

Synthetic speech detectors (SSDs) are deployed to mitigate the misuse of synthetic speech. While there are a number of advanced SSDs (Tak et al., 2021c;a; Jung et al., 2022), recent research (Müller et al., 2022; Xie et al., 2024) suggests they might struggle when facing "test domain shift". Concretely, a detector may exhibit decreased performance when presented with audio that has undergone alterations, including transcoding, playback, background noise, or even just a shift in the TTS system used. However, current research efforts are directed towards these natural changes to audio, it is important to recognize that deliberate manipulation of synthetic speech by an attacker with the intent to deceive detectors can significantly increase the likelihood of success of the attacker. However, there is a lack of systematic research on these malicious perturbations.

In this work, we conduct the first systematic study of active malicious attacks against the state-of-the-art open-source SSDs. We investigate a range of attack scenarios, considering adversaries with varying levels of access to the target SSD: those with full knowledge of the model (white-box), those who can only interact with it and observe the results (black-box), and those who cannot even query the model (agnostic); and evaluate the results using both hard-coded metrics and human ratings.

Our findings reveal that:

[*] Equal Contribution

- Increased access to the detector makes it easier for attackers to create deepfakes that can evade detection without any noticeable loss in audio quality.
- Existing open-source SSD detectors are vulnerable when facing synthetic audio generated by TTS systems never seen during training.
- VisQOL scores and human ratings show that the audio quality after attack is reasonable.
- Alarmingly, even in the agnostic setting, attackers can still bypass state-of-the-art open-source SSDs with reasonable chance.

Overall, we need more robust SSDs to mitigate the growing threat of deepfake audio misuse.

2 PRELIMINARY

2.1 TTS TECHNIQUES

TTS systems, which convert written text into speech, have a long history of development and have made remarkable progress in recent years. Early TTS systems primarily used a concatenative approach (Khan & Chitode, 2016), where speech was synthesized by joining pre-recorded speech units from a database. Despite the simplicity, they suffered from unnatural prosody and robotic-sounding speech. To address the issue, researchers proposed statistical parametric speech synthesis (Zen et al., 2009). These systems used statistical models to learn the relationship between linguistic features (e.g., phonemes, part-of-speech tags) and acoustic features (e.g., fundamental frequency, spectral envelope) and enabled more natural-sounding speech generation with improved prosody.

Recently, deep learning has revolutionized the field of TTS. Neural network-based TTS systems have surpassed traditional methods in terms of speech naturalness and intelligibility. Several architectures have been explored, including sequence-to-sequence models (Wang et al., 2017; Ping et al., 2017), attention-based models (Ren et al., 2019; 2020), and generative adversarial networks (GANs) (Kumar et al., 2019). These models can directly learn the mapping from text to speech, enabling end-to-end training and eliminating the need for complex feature engineering. Some of the popular neural TTS systems include Kreuk et al. (2022); Borsos et al. (2023); Leng et al. (2023); Saeki et al. (2023); Shen et al. (2023); Wang et al. (2023); ChatTTS (2024); Chen et al. (2024); xTTS (2024); ElevenLabs (2024); Le et al. (2024); Lux et al. (2024).

2.2 SYNTHETIC SPEECH DETECTION TECHNIQUES

In the past, detecting synthetic speech required carefully crafted features (Doddington et al., 2001; Alegre et al., 2013; Hanilçi et al., 2015; Patel & Patil, 2015; Sahidullah et al., 2015; Todisco et al., 2016). However, with the increase in available data and the development of larger models, simpler features like waveforms or spectrograms are now sufficient for effective detection. RawNet2 (Tak et al., 2021c) is a deep convolutional neural network for synthetic speech detection using merely raw waveforms. It builds upon the RawNet (Jung et al., 2020) architecture by incorporating residual connections and dilated convolutions. RawNetGATST (Tak et al., 2021a) extends RawNet2 by incorporating a graph attention network (Tak et al., 2021b) to identify key spectral or temporal features for detection. Similarly, AASIST (Jung et al., 2022) refines the graph network architecture further for improved synthetic speech detection.

3 BYPASS SYNTHETIC SPEECH DETECTION SYSTEMS

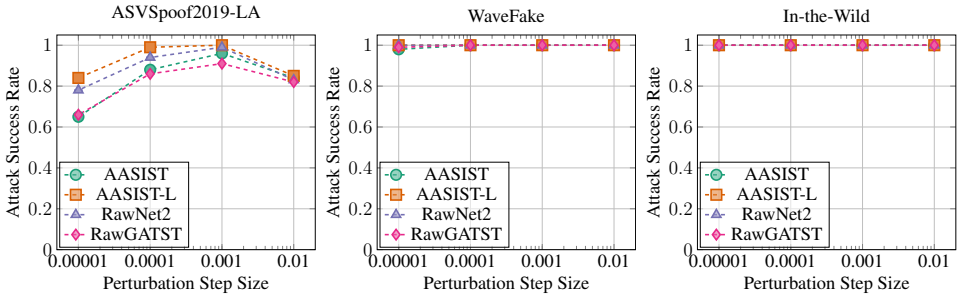
In this section, we aim to answer the following question:

Can deepfake audio be altered in ways nearly imperceptible to the human ear, but sufficient to bypass state-of-the-art detectors?

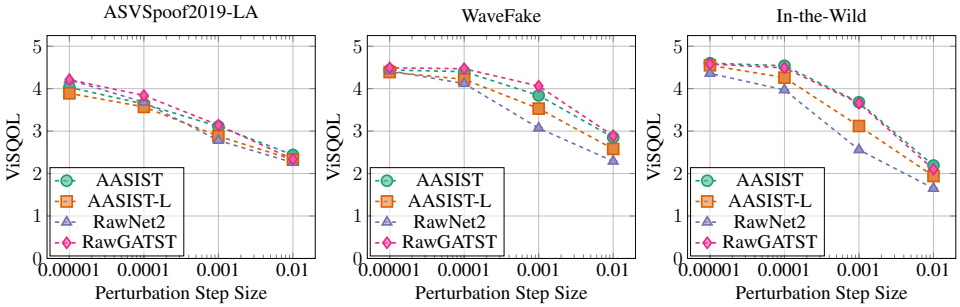
Unlike previous research that focused on natural perturbations (Müller et al., 2022; Xie et al., 2024), we consider a malicious attacker who deliberately optimizes the perturbation to evade detection. We examine this scenario under various levels of access to the detection systems, from having full knowledge (white-box), to partial knowledge (black-box), to no knowledge (agnostic).

Table 1: Baseline EERs of SSDs on the ASVSpoof2019-LA test split without attacks.

AASIST	AASIST-L	RawNet2	RawGATST
0.83%	0.99%	4.88%	3.29%



(a) Attack success rate vs. perturbation step size in PGD.



(b) ViSQOL score vs. perturbation step size in PGD.

Figure 1: Attack success rate and ViSQOL vs. perturbation step size in PGD.

3.1 Experiment Setup

We train four SOTA open-source SSDs from scratch: AASIST (Jung et al., 2022), AASIST-L (Jung et al., 2022), RawNet2 (Tak et al., 2021c) and RawGATST (Tak et al., 2021a) on ASVSpoof2019-LA train split (Todisco et al., 2019). Their equal error rates (EERs) without attacks on ASVSpoof2019-LA test split are reported in Table 1, and closely match the reported numbers in their original papers.

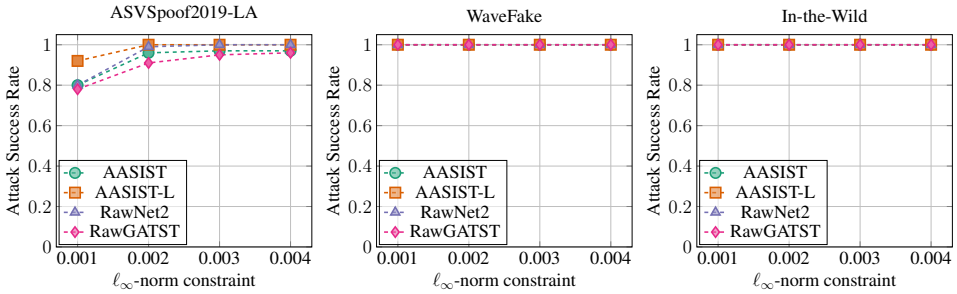
We launch attacks on three synthetic datasets: ASVSpoof2019-LA test split (Todisco et al., 2019), WaveFake (Frank & Schönherr, 2021) and In-the-wild (Müller et al., 2022). In consideration of compute resources, we randomly sub-sample 100 examples from each dataset for the attacks.

We use the attack success rate (*i.e.* the ratio of attacked examples bypassing the target detector) to measure the effectiveness of the attacks. To ensure the attack does not degrade audio quality, we use both VisQOL (Hines et al., 2015) and human ratings to confirm that the attacked audio still sounds similar to the original synthetic audio, which we refer to as "stealthiness".

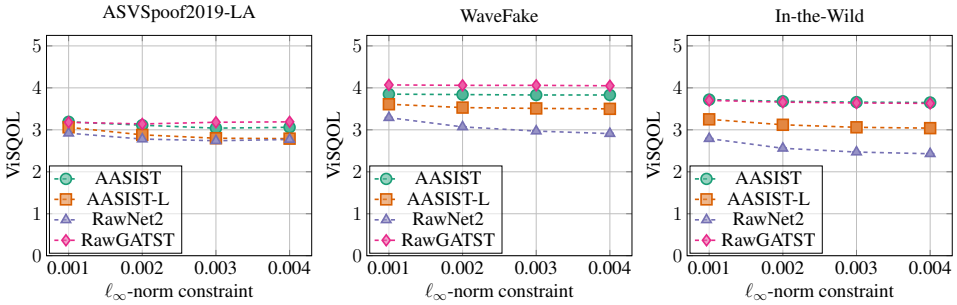
3.2 White-box Attack

Takeaway:

- Existing open-source SSDs are extremely vulnerable to white-box attacks.
- Deepfake speech from TTS not seen during training is more likely to bypass SSDs.
- White-box attacks can be highly effective and stealthy simultaneously.

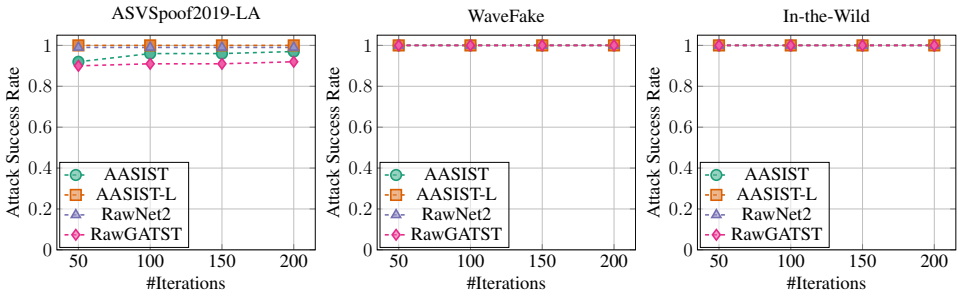


(a) Attack success rate vs. ℓ_∞-norm constraint in PGD.

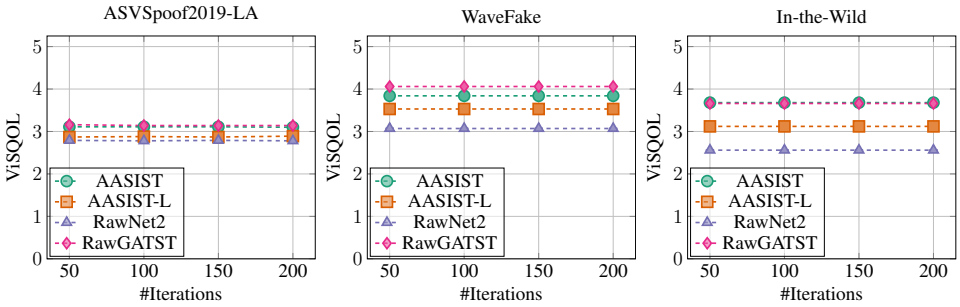


(b) ViSQOL score vs. ℓ_∞-norm constraint in PGD.

Figure 2: Attack success rate and ViSQOL vs. ℓ_∞-norm constraint in PGD.



(a) Attack success rate vs. PGD #iterations.



(b) ViSQOL score vs. PGD #iterations.

Figure 3: Attack success rate and ViSQOL vs. PGD #iterations across red-team sets.

We first study white-box attack, where the adversary has full access to the model. We choose two white-box attacks: Projected Gradient Descent (PGD) (Mkadry et al., 2017) and I-FGSM (Kurakin et al., 2018). Algorithm 1 in Appendix shows details of PGD and I-FGSM.

Projected Gradient Descent: PGD crafts adversarial examples by iteratively taking small steps in the direction that maximizes the model's error, while projecting the perturbed example back within

Table 2: Human ratings of speaker similarity between the original and PGD attacked audio.

	ASVspoof	WaveFake	In-the-wild
AASIST	0.970 ± 0.063	0.971 ± 0.046	0.985 ± 0.030
AASIST-L	0.979 ± 0.037	0.979 ± 0.045	0.975 ± 0.036
RawNet2	0.971 ± 0.077	1.000 ± 0.000	0.967 ± 0.063
RawGATST	0.997 ± 0.008	0.986 ± 0.030	0.997 ± 0.008

a certain boundary around the original input to maintain a balance between attack success rate and stealthiness.

PGD has three major hyper-parameters: perturbation step size, ℓ_∞-norm constraint, and the number of iterations. We conduct hyper-parameter search and summarize the results in Figure 1, 2, and 3.

In Figure 1, we can tell that on WaveFake and In-the-wild, the attack success rate is almost always 100% while on ASVSpoof2019-LA test the attack success rate hovers between 60% and 100% depending on the learning rate used. This reflects the fact that the detectors are more robust on test data generated by the same TTS systems as the training data (*i.e.* in-domain data), but are still vulnerable under white-box attacks with a few steps of hyper-parameter search. On the other hand, VisQOL scores keep decreasing as the perturbation step size grows. Usually, VisQOL score above 3.0 is considered reasonable quality. Thus, there exists a sweet spot of perturbation step size striking balance between attack effectiveness and stealthiness.

In Figure 2, the observation of SSDs being more robust on ASVSpoof2019-LA test holds true. However, we observe that the VisQOL scores are pretty consistent despite the changing ℓ_∞-norm constraint, which says that audio quality is insensitive to ℓ_∞-norm constraint within a certain range.

Figure 3 shows that white-box attacks are efficient, reaching maximum attack success rates and stable VisQOL scores after just 50 iterations.

We also collect human ratings on whether the PGD-attacked audio with the best hyper-parameter combination sounds like the original synthetic audio, and the results are summarized in Table 2. We can see that most human raters think the two audio sound like the same person, underscoring the potential threat of using the attacked audio for impersonation.

Iterative Fast Gradient Sign Method: I-FGSM only differs from PGD in that it only uses the sign of the gradient to perturb the input audio. It shares the same set of hyper-parameters as PGD, for which the grid search results are summarized in Figure 4, 5, and 6 and human ratings are summarized in Table 3, and the findings are similar to PGD.

3.3 Black-box Attack

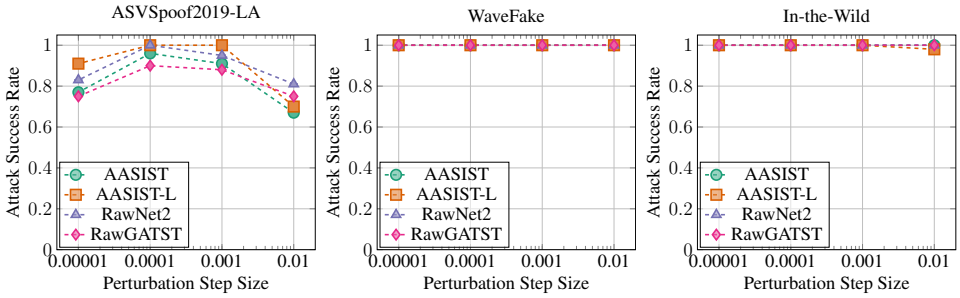
Takeaway:

- Existing open-source SSDs are still vulnerable to black-box attacks.
- Black-box attacks can be effective and stealthy simultaneously.

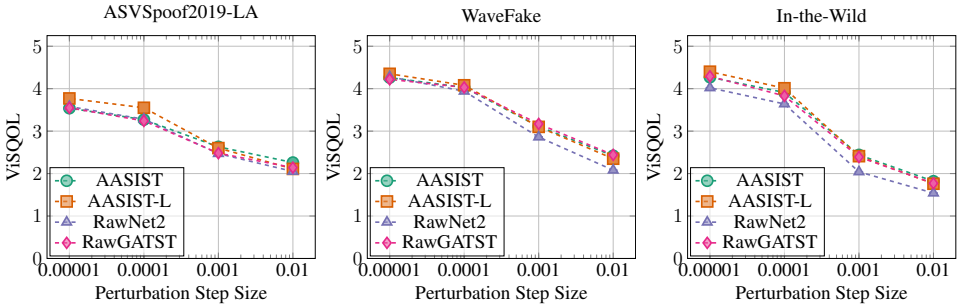
For black-box attack, we choose the Simple Black Box Attack (SimBA) (Guo et al., 2019). SimBA perturbs the input audio randomly and observes whether the prediction confidence score for "fake" class decreases or increases. If the confidence score decreases, SimBA will keep the perturbation. Otherwise, SimBA will try adding perturbation in the opposite direction and decide whether to

Table 3: Human ratings of speaker similarity between the original and I-FGSM attacked audio.

	ASVspoof	WaveFake	In-the-wild
AASIST	0.984 ± 0.020	0.960 ± 0.052	0.985 ± 0.024
AASIST-L	0.987 ± 0.022	0.986 ± 0.023	0.967 ± 0.054
RawNet2	0.980 ± 0.040	1.000 ± 0.000	0.991 ± 0.012
RawGATST	0.989 ± 0.024	0.858 ± 0.141	0.985 ± 0.024

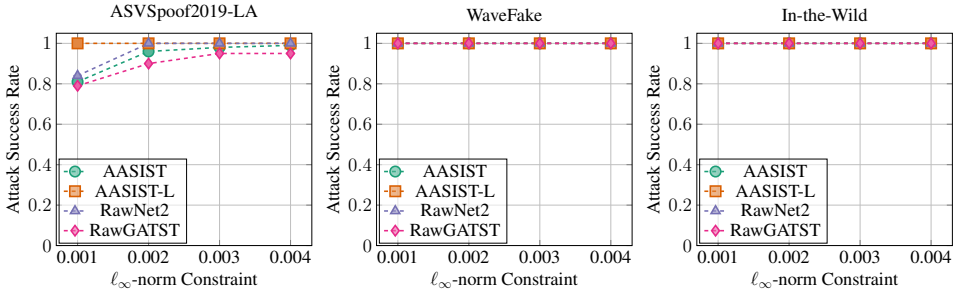


(a) Attack success rate vs. perturbation step size in I-FGSM.

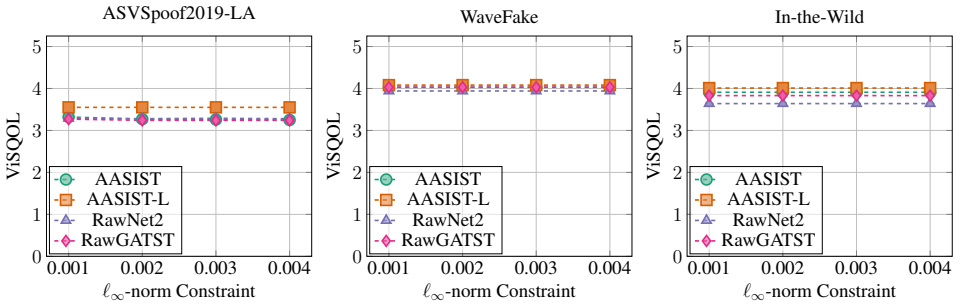


(b) ViSQOL score vs. perturbation step size in I-FGSM.

Figure 4: Attack success rate and ViSQOL vs. perturbation step size in I-FGSM.



(a) Attack success rate vs. ℓ_∞-norm constraint I-FGSM.

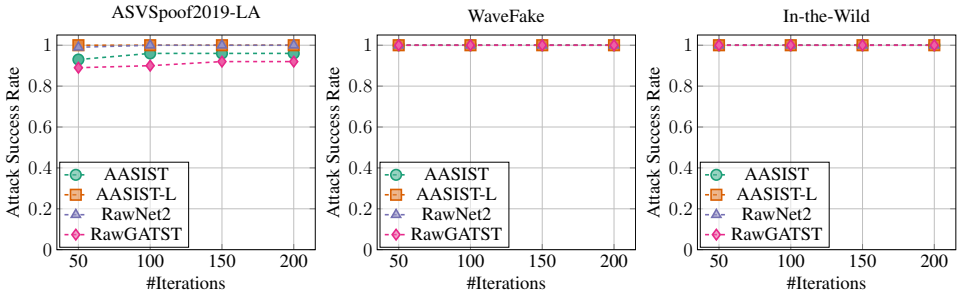


(b) ViSQOL score vs. ℓ_∞-norm constraint I-FGSM.

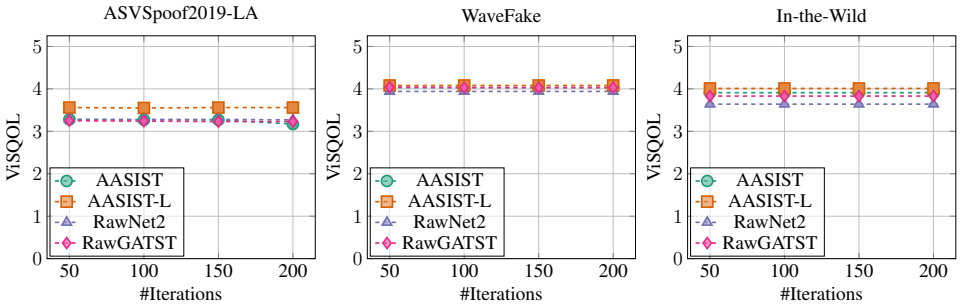
Figure 5: Attack success rate and ViSQOL vs. ℓ_∞-norm constraint in I-FGSM.

keep or discard the perturbation just as above. SimBA iteratively perturbs the input until the SSD is successfully bypassed or the budget of queries/iterations is used up. Algorithm 2 in Appendix shows the details of SimBA.

SimBA has three hyper-parameters: perturbation batch size, perturbation step size, and the number of queries. Perturbation batch size decides how many timesteps are perturbed in each query, while

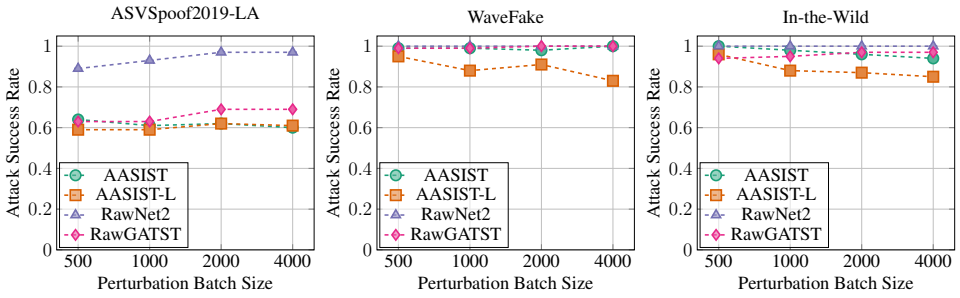


(a) Attack success rate vs. I-FGSM #iterations.

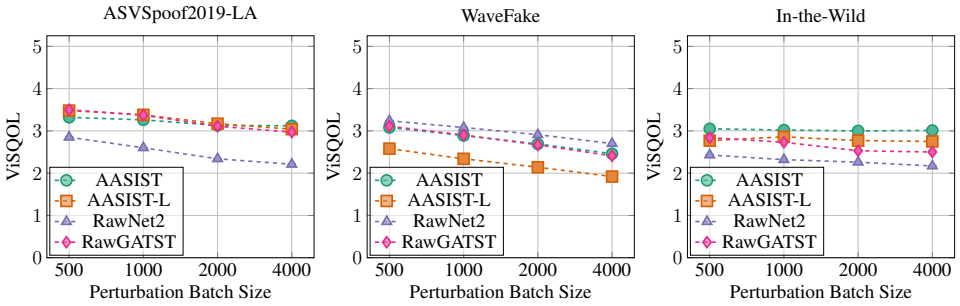


(b) ViSQOL score vs. I-FGSM #iterations.

Figure 6: Attack success rate and ViSQOL vs. I-FGSM #iterations.



(a) Attack success rate vs. SimBA perturbation batch size.

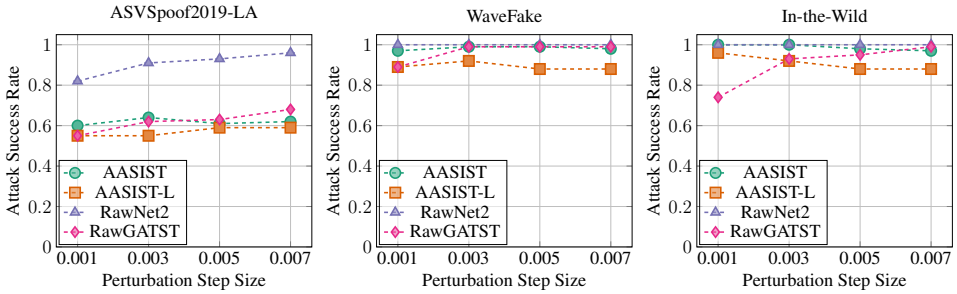


(b) ViSQOL score vs. SimBA perturbation batch size.

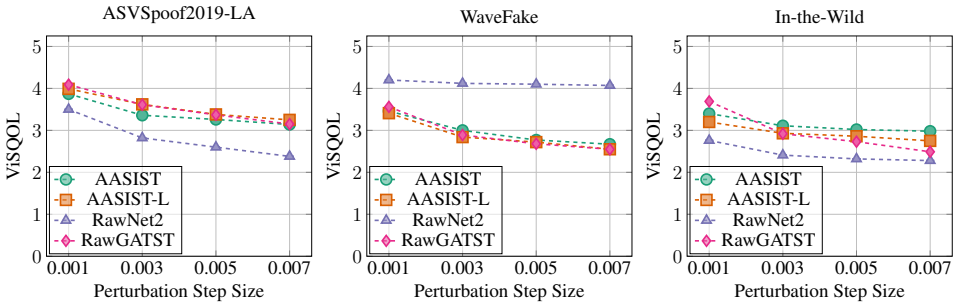
Figure 7: Attack success rate and ViSQOL vs. SimBA perturbation batch size.

perturbation step size decides which long one perturbation step on one timestep can be. The hyper-parameter search results are summarized in Figure 7, 8, and 9.

In Figure 7, we observe that on ASVSpoof2019 test, RawNet2, the least capable SSD model is still broken almost 100% but all the other 3 models are only broken 60% of all the tested examples. This draws a positive correlation between model capability and robustness. On WaveFake and In-the-

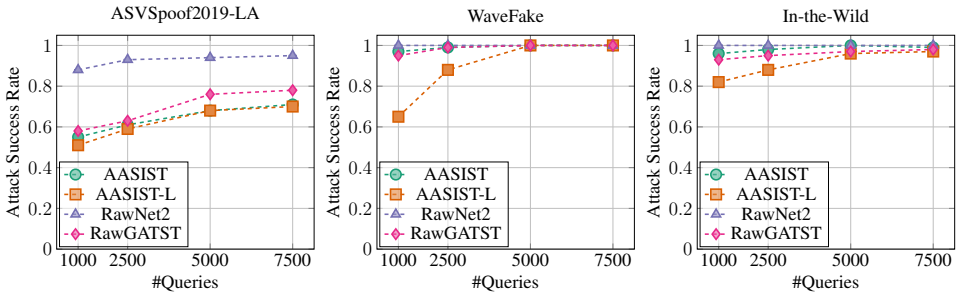


(a) Attack success rate vs. SimBA perturbation step size.

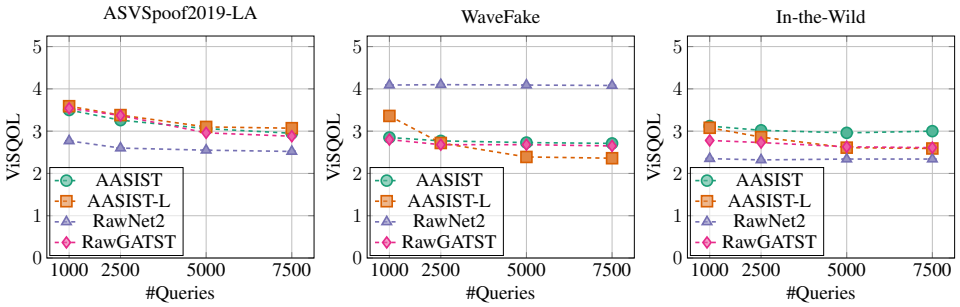


(b) ViSQOL score vs. SimBA perturbation step size.

Figure 8: Attack success rate and ViSQOL vs. SimBA perturbation step size across datasets.



(a) Attack success rate vs. SimBA #queries.



(b) ViSQOL score vs. SimBA #queries.

Figure 9: Attack success rate and ViSQOL vs. SimBA #queries across datasets.

Wild, all SSDs are broken more than 90% of the time, which confirms the previous observation that current SSD models are brittle when facing synthetic audio from TTS systems never seen during training.

Table 4: Human ratings of speaker similarity between the original and simBA attacked audio.

	ASVspoof	WaveFake	In-the-wild
AASIST	0.984 ± 0.020	0.960 ± 0.052	0.985 ± 0.024
AASIST-L	0.987 ± 0.022	0.986 ± 0.023	0.967 ± 0.054
RawNet2	0.980 ± 0.040	1.000 ± 0.000	0.991 ± 0.012
RawGATST	0.989 ± 0.024	0.858 ± 0.141	0.985 ± 0.024

Also in Figure 7, 8, and 9, we observe that ASSIST-L is the most robust model consistently, which is surprising because it's the smallest model within the 4 (See Jung et al. (2022) for the size of these models.). This observation aligns with the principle of Occam's razor, which suggests that simpler models often generalize better. A potential explanation could lie in the raggedness of the decision boundaries. Larger models, with their increased complexity, might create more intricate and potentially overfit decision boundaries. In contrast, ASSIST-L, being smaller, may form smoother decision boundaries, leading to better generalization and robustness against perturbations.

Human ratings of audio similarity is summarized in Table 4. Again the attacked audio sound highly similar to the original ones to human ears.

3.4 Agnostic Attack: Transferability of Above Attacks

Takeaway:

- For both white-box attacks and black-box attacks, transferability depends on the target model's capability on the target audio.
- Black-box attacks are more transferrable on in-domain test data than out-of-domain data.
- Transferrability of different white-box attacks are alike.

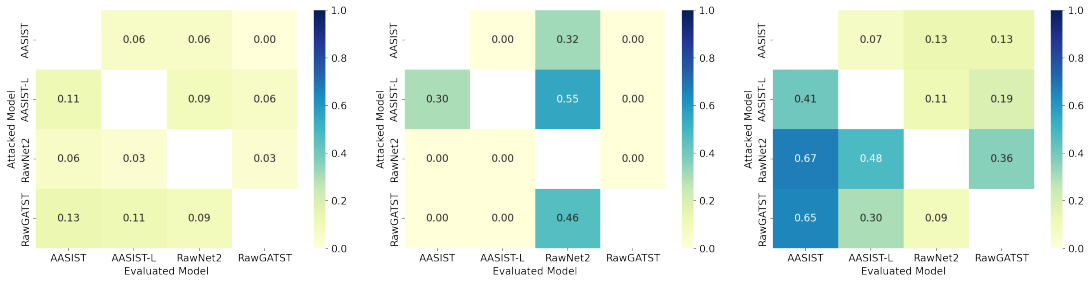
The above attacks all assume different levels of access to the SSD model which might not be accessible in practice. As a result, we want to understand whether the above attacks are transferrable: Can a successfully attacked example on one model transfer to a different model? If this is true, then the adversary can craft a proxy model themselves, attack it, and expect it to bypass the real SSD as well.

The results are summarized in Figure 10. First, we find that on out-of-domain data, some SSDs are extremely vulnerable. For example, on WaveFake, RawNet2 is extremely vulnerable under all attacks; on In-the-wild, ASSIST and AASIST-L are more vulnerable than the other two models. Second, we find that on in-domain data, black-box attacks are much more transferrable than white-box attacks. This is because 1) black-box attacks tend to add larger perturbation than white-box attacks; 2) the SSDs' decision borders are alike for in-domain data. Thirdly, we also observe high similarity between the transferrability heatmap between PGD and I-FGSM, which might be due to different white-box attacks taking gradient paths in similar directions despite small differences.

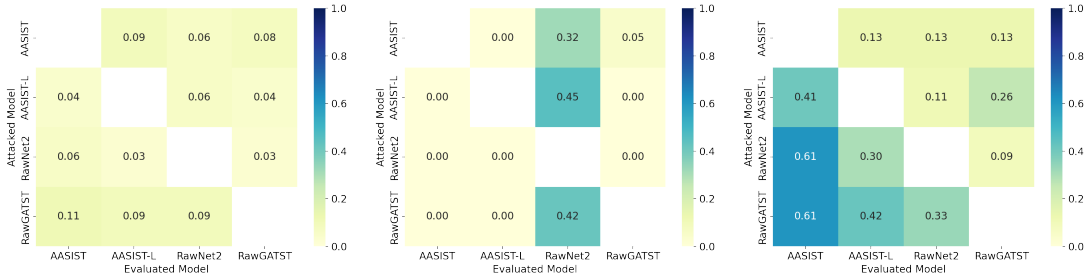
4 Conclusion

This work presents the first systematic exploration of the robustness of state-of-the-art SSDs against adversarial attacks. Our findings reveal critical implications for SSD deployment and future research.

Firstly, we demonstrate a clear correlation between system accessibility and vulnerability. Open-access SSDs, and even those with oracle access, are highly susceptible to attacks. This underscores the critical need to restrict public access to SSD models and internal workings. While complete prevention of information leakage may be challenging, measures such as rate limiting can effectively mitigate the threat of black-box attacks.



(a) PGD on ASVSpoof2019-LA (b) PGD on WaveFake (c) PGD on In-the-Wild



(d) I-FGSM on ASVSpoof2019-LA (e) I-FGSM on WaveFake (f) I-FGSM on In-the-Wild

(g) SimBA on ASVSpoof2019-LA (h) SimBA on WaveFake (i) SimBA on In-the-Wild

Figure 10: Transferability of attacks on different datasets.

Secondly, our research exposes the vulnerability of current open-source SSDs to domain shifts, including active perturbations and, notably, synthetic data generated from unseen TTS systems. This highlights two key recommendations:

- Comprehensive and Continuously Updated Training Data: SSD training datasets should encompass a diverse range of TTS systems and be regularly updated to incorporate new systems, ensuring broad coverage against evolving spoofing techniques.
- Data Composition Confidentiality: Maintaining the confidentiality of training data composition is crucial to prevent attackers from exploiting this knowledge for targeted attacks, as evidenced by the significant advantage gained from TTS system shifts even in agnostic attacks.

Our findings also draw a parallel between the interplay of SSD and TTS development and an offline GAN, suggesting a future where TTS systems may achieve near-perfect human mimicry. This emphasizes the crucial need for complementary audio authenticity techniques, such as robust watermarking (Liu et al., 2024), to bolster the security of SSD systems.

In conclusion, this study serves as a critical analysis of the current state of open-source SSD robustness. By exposing key vulnerabilities and providing actionable recommendations, we aim to guide future research and development efforts towards building more secure and resilient SSD systems in an ever-evolving landscape of speech synthesis and spoofing technologies.

REFERENCES

Federico Alegre, Ravichander Vipperla, Asmaa Amehraye, and Nicholas Evans. A new speaker verification spoofing countermeasure based on local binary patterns. In *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon: France (2013)*, pp. 5p, 2013.

Fadi Biadsy, Youzheng Chen, Isaac Elias, Kyle Kastner, Gary Wang, Andrew Rosenberg, and Bhuvana Ramabhadran. Zero-shot cross-lingual voice transfer for tts. *arXiv preprint arXiv:2409.13910*, 2024.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pp. 2709–2720. PMLR, 2022.

ChatTTS. Chattts - text-to-speech for conversational scenarios. https://chattts.com/, 2024. Accessed: 07/16/2024.

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*, 2024.

Devin Coldewey. Six million fine for robocaller who used ai to clone biden's voice. https://techcrunch.com/2024/05/23/6m-fine-for-robocaller-who-used-ai-to-clone-bidens-voice, 2024. Online; accessed 29 May 2024.

Erica Cooper, Cheng-I Lai, Yusuke Yasuda, Fuming Fang, Xin Wang, Nanxin Chen, and Junichi Yamagishi. Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6184–6188. IEEE, 2020.

George R Doddington et al. Speaker recognition based on idiolectal differences between speakers. In *Interspeech*, pp. 2521–2524, 2001.

ElevenLabs. High quality, human-like ai voice generator. `https://elevenlabs.io/text-to-speech`, 2024. Accessed: 07/16/2024.

Joel Frank and Lea Schönherr. Wavefake: A data set to facilitate audio deepfake detection. *arXiv preprint arXiv:2111.02813*, 2021.

Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple black-box adversarial attacks. In *International conference on machine learning*, pp. 2484–2493. PMLR, 2019.

Cemal Hanilçi, Tomi Kinnunen, Md Sahidullah, and Aleksandr Sizov. Classifiers for synthetic speech detection: A comparison. 2015.

Andrew Hines, Jan Skoglund, Anil C Kokaram, and Naomi Harte. Visqol: an objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015:1–18, 2015.

Jee-weon Jung, Seung-bin Kim, Hye-jin Shim, Ju-ho Kim, and Ha-Jin Yu. Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms. *arXiv preprint arXiv:2004.00526*, 2020.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 6367–6371. IEEE, 2022.

Rubeena A Khan and Janardan Shrawan Chitode. Concatenative speech synthesis: A review. *International Journal of Computer Applications*, 136(3):1–6, 2016.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022.

Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32, 2019.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pp. 99–112. 2018.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36, 2024.

Yichong Leng, Zhifang Guo, Kai Shen, Xu Tan, Zeqian Ju, Yanqing Liu, Yufei Liu, Dongchao Yang, Leying Zhang, Kaitao Song, et al. Prompttts 2: Describing and generating voices with text prompt. *arXiv preprint arXiv:2309.02285*, 2023.

Hongbin Liu, Moyang Guo, Zhengyuan Jiang, Lun Wang, and Neil Zhenqiang Gong. Audiomarkbench: Benchmarking robustness of audio watermarking. *arXiv preprint arXiv:2406.06979*, 2024.

Florian Lux, Sarina Meyer, Lyonel Behringer, Frank Zalkow, Phat Do, Matt Coler, Emanuël A. P. Habets, and Ngoc Thang Vu. Meta Learning Text-to-Speech Synthesis in over 7000 Languages. In *Interspeech*. ISCA, 2024.

Aleksander Mkadry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *stat*, 1050(9), 2017.

Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? *arXiv preprint arXiv:2203.16263*, 2022.

Tanvina B Patel and Hemant A Patil. Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech. In *Interspeech*, pp. 2062–2066, 2015.

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv preprint arXiv:1710.07654*, 2017.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32, 2019.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

Takaaki Saeki, Heiga Zen, Zhehuai Chen, Nobuyuki Morioka, Gary Wang, Yu Zhang, Ankur Bapna, Andrew Rosenberg, and Bhuvana Ramabhadran. Virtuoso: Massive multilingual speech-text joint semi-supervised learning for text-to-speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Md Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. A comparison of features for synthetic speech detection. 2015.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*, 2023.

Hemlata Tak, Jee-weon Jung, Jose Patino, Madhu Kamble, Massimiliano Todisco, and Nicholas Evans. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. *arXiv preprint arXiv:2107.12710*, 2021a.

Hemlata Tak, Jee-weon Jung, Jose Patino, Massimiliano Todisco, and Nicholas Evans. Graph attention networks for anti-spoofing. *arXiv preprint arXiv:2104.03654*, 2021b.

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6369–6373. IEEE, 2021c.

Massimiliano Todisco, Héctor Delgado, and Nicholas WD Evans. A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients. In *Odyssey*, volume 2016, pp. 283–290, 2016.

Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. Asvspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*, 2019.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

Yuankun Xie, Yi Lu, Ruibo Fu, Zhengqi Wen, Zhiyong Wang, Jianhua Tao, Xin Qi, Xiaopeng Wang, Yukun Liu, Haonan Cheng, et al. The codecfake dataset and countermeasures for the universally detection of deepfake audio. *arXiv preprint arXiv:2405.04880*, 2024.

xTTS. High quality, human-like ai voice generator. https://github.com/coqui-ai/TTS, 2024. Accessed: 07/16/2024.

Zhen Ye, Zeqian Ju, Haohe Liu, Xu Tan, Jianyi Chen, Yiwen Lu, Peiwen Sun, Jiahao Pan, Weizhen Bian, Shulin He, et al. Flashspeech: Efficient zero-shot speech synthesis. *arXiv preprint arXiv:2404.14700*, 2024.

Heiga Zen, Keiichi Tokuda, and Alan W Black. Statistical parametric speech synthesis. *speech communication*, 51(11):1039–1064, 2009.

Algorithm 1 White-box Attacks: PGD and I-FGSM

Input: Waveform audio $s \in \mathbb{R}^T$, SSD model f, perturbation step size α, maximum number of iterations T, real audio class $\mathcal{R}$, attack $A \in \{$'PGD', 'I-FGSM'$\}$, ℓ_∞-norm constraint ϵ
Output: Adversarial perturbation $\delta \in \mathbb{R}^T$

$\delta \leftarrow 0$ ▷ initialization
for $t \leftarrow 1$ to T **do**
  if A == 'PGD' **then**
    $\delta \leftarrow \delta - \alpha \cdot \nabla_\delta l_{CE}(f(s+\delta), \mathcal{R})$ ▷ gradient descent with cross entropy loss l_{CE}
  end if
  if A == 'I-FGSM' **then**
    $\delta \leftarrow \delta - \alpha \cdot \text{sign}(\nabla_\delta l_{CE}(f(s+\delta), \mathcal{R}))$ ▷ FGSM with cross entropy loss l_{CE}
  end if
  $\delta \leftarrow \text{clip}(\delta, -\epsilon, \epsilon)$ ▷ projection
  if $f(s+\delta) == \mathcal{R}$ **then** ▷ attack success
    Break
  end if
end for
return δ

Algorithm 2 Black-box Attack: SimBA

Input: Waveform audio $s \in \mathbb{R}^T$, SSD model f, perturbation step size α, perturbation batch size q, maximum number of queries Q, real audio class $\mathcal{R}$, ℓ_∞-norm constraint ϵ
Output: Adversarial perturbation $\delta \in \mathbb{R}^T$

$\delta \leftarrow 0, t \leftarrow 0$ ▷ initialization
$p \leftarrow f(s, \mathcal{R}), t \leftarrow t+1$ ▷ initializing highest probability to predict real audio class
while $t < T$ **do**
  if $f(s+\delta) == \mathcal{R}$ **then** ▷ attack success
    Break
  end if
  $r \in \mathbb{R}^T$ and $r \leftarrow 0$
  Randomly choose q dimensions from r without replacement
  Randomly add α or $-\alpha$ to the chosen q dimensions in r
  $t \leftarrow t+1$ ▷ one more query for f below
  if $f(s+\delta+r, \mathcal{R}) > p$ **then**
    $p \leftarrow f(s+\delta+r, \mathcal{R})$ ▷ update highest probability to predict real audio class
    Continue
  else
    $t \leftarrow t+1$ ▷ one more query for f below
    if $f(s+\delta-r, \mathcal{R}) > p$ **then**
      $p \leftarrow f(s+\delta-r, \mathcal{R})$ ▷ update highest probability to predict real audio class
      Continue
    end if
  end if
end while
return δ